Jiang TB, Xia GS, Lu QK *et al.* Retrieving aerial scene images with learned deep image-sketch features. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 32(4): 726–737 July 2017. DOI 10.1007/s11390-017-1754-7

Retrieving Aerial Scene Images with Learned Deep Image-Sketch Features

Tian-Bi Jiang [1,2], Gui-Song Xia [1], *Senior Member, IEEE, Member, CCF*, Qi-Kai Lu [2,*], and Wei-Ming Shen [1]

[1] *State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University Wuhan 430079, China*

[2] *School of Electronic Information, Wuhan University, Wuhan 430072, China*

E-mail: {jiangtianbi, guisong.xia, qikai_lu, shenwm}@whu.edu.cn

Received December 20, 2016; revised May 26, 2017.

Abstract This paper investigates the problem of retrieving aerial scene images by using semantic sketches, since the state-of-the-art retrieval systems turn out to be invalid when there is no exemplar query aerial image available. However, due to the complex surface structures and huge variations of resolutions of aerial images, it is very challenging to retrieve aerial images with sketches and few studies have been devoted to this task. In this article, for the first time to our knowledge, we propose a framework to bridge the gap between sketches and aerial images. First, an aerial sketch-image database is collected, and the images and sketches it contains are augmented to various levels of details. We then train a multi-scale deep model by the new dataset. The fully-connected layers of the network in each scale are finally connected and used as cross-domain features, and the Euclidean distance is used to measure the cross-domain similarity between aerial images and sketches. Experiments on several commonly used aerial image datasets demonstrate the superiority of the proposed method compared with the traditional approaches.

Keywords sketch, aerial image retrieval, multi scale, deep cross-domain model

1 Introduction

The rapid development of satellite imaging sensors brings us huge amounts of high-quality aerial images, which are widely used in environmental monitoring and geological survey[1-3]. It is imperative to browse the remote sensing images we need efficiently and effectively[4]. Aerial image retrieval is usually implemented by matching keywords or setting an exemplar image as a query[5-7]. Recently, content-based aerial image retrieval has been systematically investigated, but the previous studies mostly rely on low-level features or combing them with a better representation method[8-10]. For instance, Liu *et al.*[11] proposed a region-level semantic mining approach which is still based on the hand-crafted features.

The existing methods require users to obtain an aerial image as input exemplar when they want to search for target scene images. However, if we intend to search an interested object as an airport but no airport image is available at hand, the traditional approaches would lose efficiency. Considering that the sketch is an abstract expression and can be depicted as an interested object, we simply draw a sketch and use it to retrieve the required aerial images to solve this problem.

It is recognized that sketches have been used to depict our visual world since prehistoric times[12-13]. More specifically, with the popularization of touch screen, drawing an abstract sketch to retrieve the objective image becomes practical. Some studies have been done to retrieve natural images using sketches[14-16]. For instance, Eitz *et al.*[17] and Hu and Collomosse[18] both employed modified histogram-of-gradient (HOG) descriptors coupled with mid-level encoding like bag-of-words to query natural images with sketches. Cao *et*

Regular Paper

Special Issue on Deep Learning

This work was supported by the National Natural Science Foundation of China under Grant Nos. 41501462 and 91338113.

*Corresponding Author

al.[19] proposed an edgel structure and used it to measure sketch-image pair similarity by indexable oriented chamfer matching. Chen *et al.*[20] proposed a method for generating photo-realistic pictures from a casually drawn sketch with added text labels. Convolutional neural networks are used in [21-22] to accomplish sketch-based image retrieval and perform well. Wang *et al.*[23] collected images corresponding to TU-Berlin sketch dataset[12] and constructed a sketch-image dataset HUST-SI, which is used to fine-tune the Alexnet, and a model that can retrieve natural images with sketches was obtained. In a word, it has been demonstrated that cross-domain image-sketch comparison can improve the image retrieval performance when no exemplar query image is available. Although the sketch has been successfully applied in the natural image retrieval, few studies have been devoted to sketch-based aerial image retrieval. Owing to the complex surface structures and huge variations of image resolutions, it is very challenging to measure the similarity between such a simple sketch and a fairly complex aerial image. The ambiguity inherent in sketches and the gap between aerial images and sketches bring a great difficulty to sketch-based aerial image retrieval. The existing methods developed on natural images lose their efficacy when it comes to aerial images. Thus, it is of great interest to investigate the retrieval of aerial scene images using semantic sketches.

In this paper, we achieve this sketch-based aerial image retrieval task on the basis of a multi-scale deep cross-domain image representation model. Moreover, we contribute an aerial sketch-image database Aerial-SI. Specifically, the natural sketch-image dataset is adopted to fine-tune Alexnet, and a preliminary model is obtained. Then, we augment the sketches and images in Aerial-SI, and build a deep model by retraining the preliminary model with the augmented data. To represent the detail of various sketches, we introduce a multi-scale deep model. The multi-scale strategy is implemented by resizing images and sketches with various levels of details and training each scale as described above with these augmented data. The fully-connected layers are set as cross-domain features, and we connect features of each scale in the network to obtain a multi-scale deep cross-domain image representation. Furthermore, to measure the similarity between sketches and images, Euclidean distance between their cross-domain features is employed. Experiments conducted on the UCM dataset[24] and the RS19 dataset[25] demonstrate that our method shows satisfactory performance.

In the following parts of the paper, we present our multi-scale deep model for cross-domain feature extraction in Section 2, and then we describe how we collect aerial sketch-image dataset in Section 3. The experimental results and further analysis are discussed in Section 4, and finally this paper is concluded in Section 5.

2 Methodology

2.1 General Idea

Given a query sketch S and a set of candidate aerial images $\mathcal{I} = \{I_n\}_{n=1,\ldots,N}$, we want to compute the cross-domain similarity between S and $I_n \in \mathcal{I}$ in order to rank the images in dataset $\mathcal{I}$. To rank the true matches at the top, we need to obtain efficient cross-domain features and pursue a good similarity measure.

Inspired by the work in [23], which reported that deep features learned by the Alexnet with annotated sketch-image pairs can efficiently encode the cross-domain visual information and handle the sketch-based image retrieval, we build a deep network to compute cross-domain features for aerial image retrieval. In addition, to depict the complex surface structures in aerial images, we propose a multi-scale strategy to aggregate features with different levels of details.

We use a natural sketch-image dataset, in which the corresponding sketches and images are regarded as the same category, to retrain Alexnet and construct a preliminary model. Sketches are rotated to multi-angle ones, and both sketches and images are flipped to augment the dataset. We contribute an aerial sketch-image dataset with sketches and images in the same category. Considering the complexity and diversity of aerial images, we augment the aerial sketch-image dataset by rotating and flipping. We rotate sketches by calculating coordinates to preserve effective information. The aerial dataset is then resized to specific size to fine-tune the preliminary model. Through the above work, a single aerial deep model is obtained.

To overcome the ambiguity inherent in sketches and the gap between sketches and aerial images, we build a multi-scale deep model which is trained on the datasets containing different levels of details. The sketches of natural and aerial sketch-image datasets are processed to various levels of details. The images in sketch-image datasets are also blurred to multi-scale with their edges preserved. Both natural and aerial sketch-image datasets are processed to four scales with the same

multi-scale parameters. For each scale, sketches and images are also corresponding, which means sketches with less details correspond to images blurred to a larger scale with less texture. We use these multi-scale image-sketch datasets to train model at each scale. The 7th fully-connected layers of each scale are extracted and connected as cross-domain features, and the Euclidean distance of features is employed to retrieve aerial scene images. The overview flow chart of our method is shown in Fig.1, and the details are explained as follows.

2.2 Data Augmentation

Since the size of aerial database is too small to retrain a convolutional neural network (CNN) model and the input of Alexnet is fixed, data augmentation is necessary to reduce over-fitting and we need to resize the input to a fixed size. Therefore, aerial images are resized to 256×256 using bilinear interpolation, and then we sample nine patches with the size of 227×227 and flip the patches horizontally. In this way, the size of the database turns out to be 18 times of the original one.

We perform data augmentation by horizontally reflecting and rotating (in the range of $[-45°, +45°]$) the sketches. Without brightness and texture, sketches cannot be effectively resized using typical image resizing method. In order to preserve the sketches in transformations, the coordinates of sketch points are calculated. The reflection and the rotation are achieved by coordinate calculation, and we obtain the augmented sketches in 227×227 after the above operation. Thus the size of sketch database is also increased to 18 times of the original one.

2.3 Training Cross-Domain Model

Using a sketch to search for a similar aerial image is a challenging task for the difficulty to bridge the gap between sketches and images. Moreover, the aerial images which contain complex and various surface features bring more difficulties to this task. We want to obtain cross-domain features that are domain-invariant and can obtain the common characteristics of sketches and images. To achieve this, for each scale of the network, a deep cross-domain model is trained with the sketch-image dataset. Details of the training strategy are as follows.

The Alexnet[26] trained on the ImageNet dataset is a powerful image classifier. It is able to obtain the strong image features, and thus we use the natural sketch-image dataset to retrain Alexnet and obtain a preliminary model. Since Alexnet is trained by millions of natural images and the natural sketch-image dataset is much more massive than the aerial image-sketch dataset, the preliminary model can provide a good initial value to speed up the training and gain a good result at the same time.

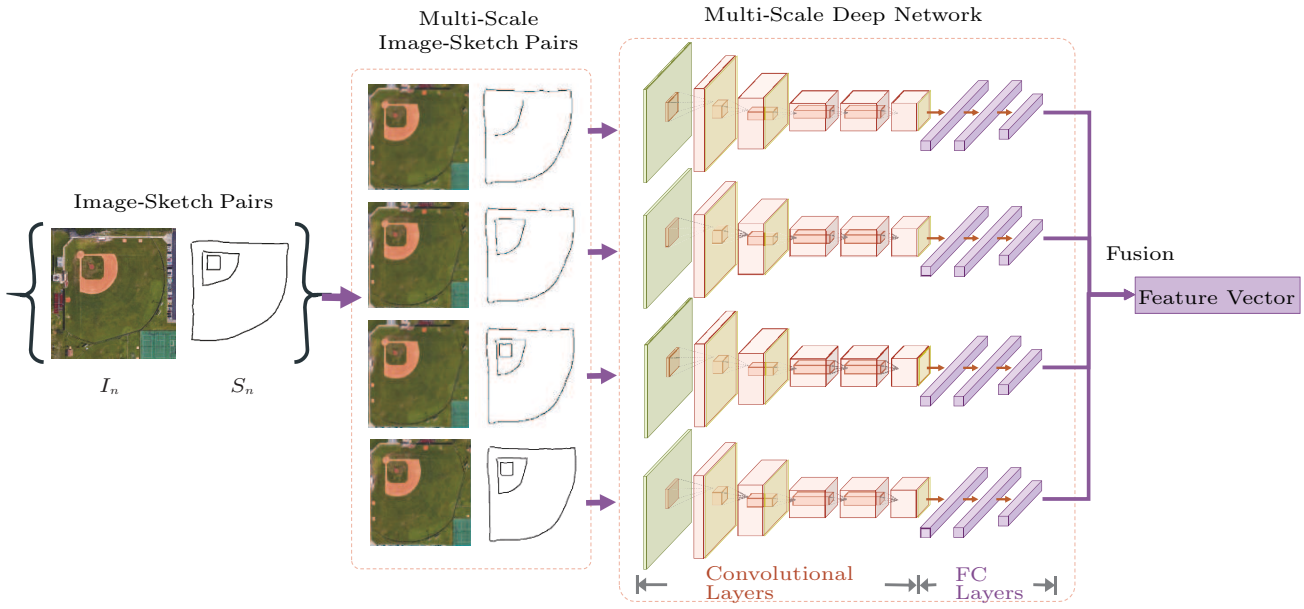


Fig.1. Learning a multi-scale aerial image representation network: the image-sketch dataset with various levels of details is used to train a multi-scale deep network. The outputs of the network are cross-domain visual features concatenated by the fully-connected layer of each scale.

The aerial images are different from the natural images on visual angel, spatial structure, and are generally with larger inter-class difference. These differences enhance the difficulty to retrieve aerial images with a free-hand sketch. We try to narrow the gap between aerial scene images and semantic sketches by fine-tuning the preliminary model. The mixing aerial sketch-image dataset is equally divided into three parts. Two parts serve as the training set and the remaining part as the validation set. Then we set the number of outputs of the network to adapt to the number of categories of dataset. For the limitation of data size, we only fine-tune the net from the conv5 layer to prevent over-fitting. As shown in Fig.2, the modified Alexnet in preliminary model contains five convolutional layers and three fully-connected layers (FC, FC6, and FC7). So far, the single deep aerial sketch-image model showed in Fig.2 is obtained to extract the cross-domain features.

2.4 Building a Multi-Scale Network

We have used the natural sketch-image dataset to retrain Alexnet with the method in [23] and obtain the preliminary model. The aerial sketch-image dataset is used to train a single aerial cross-domain model in the previous work. The sketch is a quite abstract expression and has inherent ambiguity, and it brings fairly diversity when it comes to the same category. Thus we build a multi-scale network to overcome this problem.

As the regular pattern is to draw the main outline first when sketching an object, we rank the strokes of sketches according to the length of strokes. Then we take the top 20%, 40% and 80% to get sketches with different levels of details. The original sketch is regarded as the most detailed sketch, and we obtain multi-scale sketches with four levels of details. Correspondingly, we use the rolling guidance filter (RGF)[27] to get different fuzzy scales of edge-preserving images. Thus we obtain images with varying details and mixing them with multi-scale sketches. In each scale, the model is trained as Subsection 2.3 describes and we finally build a multi-scale network that can obtain different levels of details.

2.5 Sketch-Based Aerial Image Retrieval

Traditional hand-crafted features lose their efficacy on sketch-based aerial image retrieval due to the complex surface structures and the huge variations of resolutions of aerial images. These methods cannot extract cross-domain features to represent both sketches and images well.

The output of the multi-scale deep sketch-image model is the probability of each class the input image belongs to. Since the model is trained by mixing data, the front layers could be treated as cross-domain features. We remove the last softmax layer of the deep aerial sketch-image model and get the last and the penultimate fully-connected layer. The 6th and the 7th fully-connected layer output with 4 096 dimensions are employed as cross-domain features, or they can be connected as 8 192-dimensional features. We simply connect the output of each scale to get the multi-scale cross-domain features. Thus the similarity of sketches and aerial images can be simply measured by Euclidean distance. Aerial images in datasets are ranked according to the similarity, and the top ones are set as retrieval results. The retrieval procedure is shown in Fig.3. In

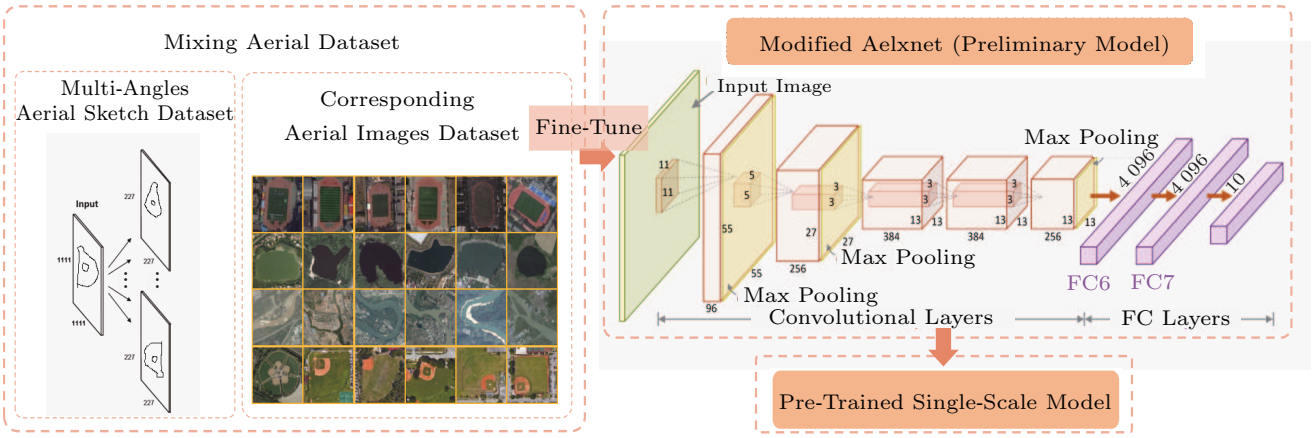


Fig.2. Training cross-domain model for each scale: the natural sketch-image dataset has been used to train a preliminary model, and the aerial sketch-image dataset with corresponding scale is augmented to fine-tune the preliminary model and get the single cross-domain model.

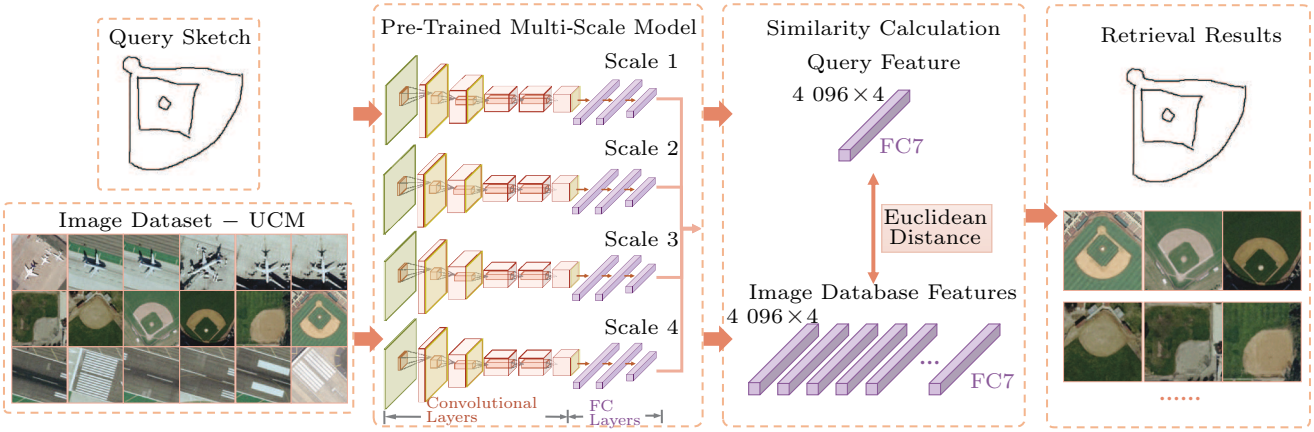


Fig.3. Retrieving aerial image with a sketch: both the sketch and the image can be represented by cross-domain features which are extracted from the multi-scale cross-domain model. The similarity of the sketch and the image is measured by Euclidean distance and the top similar images are found in this way.

this way, we achieve the goal of retrieving aerial images by using a free-hand semantic sketch.

3 Dataset

In this work, we contribute an aerial sketch-image dataset Aerial-SI with 10 categories, and each category contains aerial images from AID dataset[28-29] and sketches we collected corresponding to them. The corresponding sketches and aerial images are regarded as the same category, and thus we obtain an aerial sketch-image dataset Aerial-SI. We select 10 categories which contain salient object that can be sketched from AID dataset. The selected categories are airport, baseball field, beach, bridge, playground, pond, river, stadium, storage tanks and viaduct. The collected dataset comprises of 3 300 high-resolution aerial images.

To guarantee the sketch-based aerial image retrieval system is available for ordinary persons, 25 volunteers who are amateur in painting are recruited to collect the semantic sketches. Specifically, 15 of the volunteers are experts in remote sensing while the others are not. The instruction shows three random images out of each class, and the volunteers draw two sketches to describe what they just saw by using a touchscreen device. Then, 40 simple and clear sketches are chosen for each category, and thus, a semantic remote sensing sketch dataset of 400 sketches is obtained. Some samples of Aerial-SI are showed in Fig.4.

HUST-SI is a natural sketch-image dataset, and the corresponding sketch and image classes in the dataset are regarded as the same categories. It has 20 cate-

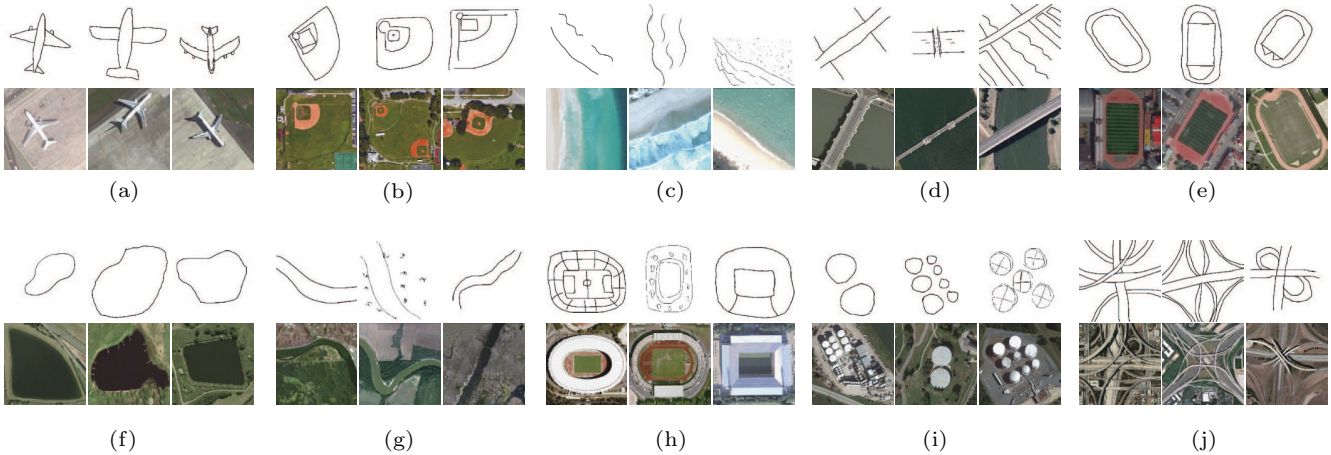


Fig.4. Samples of Aerial-SI: some corresponding sketch-image pairs are shown for each class in Aerial-SI. There are totally 10 classes containing 400 sketches and 3 300 aerial images. (a) Airport. (b) Baseball field. (c) Beach. (d) Bridge. (e) Play ground. (f) Pond. (g) River. (h) Stadium. (i) Storage tanks. (j) Viaduct.

gories with 20 000 sketches and 31 824 images. We use the above two training datasets to train our deep model. To evaluate the effectiveness of the proposed method, the experiments are performed on two manually labeled datasets, UCM and RS19. UCM[24] is a remote sensing image dataset with 21 land-use classes, in which each category contains 100 high spatial resolution images. RS19[25] is a remote sensing image dataset with 19 classes, which contains 50 images for each class.

4 Experimental Results

4.1 Experimental Settings

4.1.1 Parameter Settings

The edgelink tool is used to get the strokes of sketches. The strokes are sequenced according to the length to synthesise variously detailed sketches by preserving the top 20%, 40% and 80%. The original sketch is set as the most detailed sketch and we get sketches with four scales.

For images, we use the rolling guidance filter to obtain images with various details. The iteration of RGF is set to 4 and the spatial scale parameter σ_s varies from 3, 4, to 5 to obtain the multi-scale images. Edges vary corresponding to the images.

For each scale, we fine-tune the preliminary model from the 5th convolution layer. Based on the pre-trained model, the learning rate is set to a smaller value (0.001) in the experiments. The other parameters are the same with Alexnet and we finally get a cross-domain model trained by 160 000 times of iteration.

4.1.2 Data Mixing and Pre-Processing

Due to the small data size of sketches and the similarity between sketches and edge maps, we design two data mixing strategies: 1) aerial images and sketches (AS): the corresponding aerial images and sketches are mixed to obtain dataset Aerial-SI, which contains 3 700 images; 2) aerial images, sketches and edge maps (ASE): the edge detector in [30] is utilized to get the edge maps of aerial images. We set the corresponding remote sensing images, sketches and edge maps as the same category and obtain a dataset with 7 000 images.

The aerial images, sketches, edge maps are all augmented as Subsection 2.2 describes. Thus we obtain $400 \times 9 \times 2 = 7\,200$ semantic sketches, $3\,300 \times 9 \times 2 = 59\,400$ aerial images and 59 400 edge maps. The mixing datasets are randomly divided into three groups, and two thirds of them are set as the training set while the rest is the validation set.

4.2 Baselines

To test the effectiveness of our approach, we compare our model with several hand-craft feature baselines.

GIST. We use GIST[31] descriptor to represent aerial images and query sketches, and compare their similarity by Euclidean distance.

BoW. Since bag-of-words (BoW) is an effective scene descriptor, we consider a common approach of generating a BoW descriptor with Dense-SIFT[32] to represent images and sketches. The Dense-SIFT is extracted with 16×16 pixel patches computed over a grid with spacing of 8 pixels. Then the features are compared using histogram intersection pyramid matching kernels[33].

SIFT+SPM. We employ spatial pyramid[34] with three levels using Dense-SIFT descriptors of which the parameters are set as the above in BoW. We use the vocabulary of 200 visual words and thus images and sketches are represented with 4 200 dimensional features. The similarity is measured by histogram intersection.

GF-HOG. Histogram of oriented gradients (HOG)[35] features are popular and powerful for sketch-recognition and sketch-based image retrieval. We use gradient field HOG descriptor (GF-HOG)[18] with BoW to represent images and sketches. The GF-HOG descriptor represents image edges and sketches using a dense gradient field, and the codebook is set to 1 000. The similarity is compared using histogram intersection.

HED+GF-HOG. We use holistically-nested edge detection (HED)[36] to obtain edges of aerial images, and then compare edges and sketches to bridge the gap between images and sketches. The GF-HOG descriptor of which parameters are set as the above is employed to describe the images. The similarity between aerial images and sketches is measured using histogram intersection.

GoogLeNet. GoogLeNet is used in our method with a single-scale strategy to validate the generality of our method. The natural sketch-image dataset is used to retrain the whole GoogLeNet as described above in our method, and then the aerial sketch-image dataset is employed to fine-tune the network from the inception5b layer. The average pool layer is employed as the cross-domain feature. The similarity is measured using Euclidean distance.

4.3 Results

We evaluate our method on UCM and RS19. The aerial images without an object are difficult to sketch, for instance, some of the aerial scene classes like meadow, forest are almost represented by texture information, and others like port, parking do not have a salient object. Therefore, we only retrieve the remote sensing images which contain a salient object. We select 11 categories from the UCM dataset and six categories from RS19 to depict and retrieve these categories from the whole dataset. We collect five query sketches for each class with a salient object to evaluate the cross-domain model.

When using the output of the 7th fully-connected layer as a cross-domain feature to retrieve images with query sketch, we simply calculate the Euclidean distance to measure the similarity. The two data mixing strategies described above are compared on UCM and RS19 using single-scale deep model. Fig.5 shows the top K precision of two strategies tested on different datasets. The results demonstrate that ASE strategy always performs better than AS in this task, and thus the experiments are carried out with ASE data mixing strategy in the following parts.

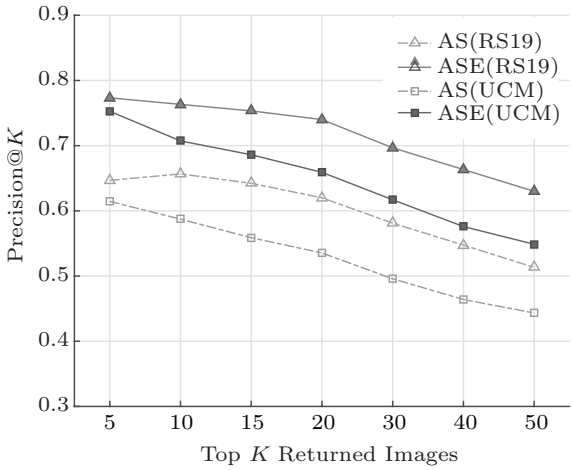
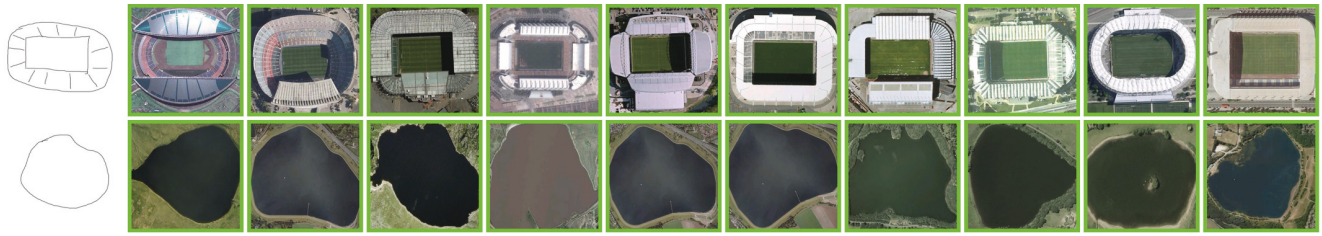


Fig.5. Top K precision of two data mixing strategies: AS, ASE.

The top 10 retrieval results of typical classes in UCM and RS19 using multi-scale deep model are showed in Fig.6. The green box means the retrieval

(a)



(b)

Fig.6. Retrieval results using multi-scale deep model: the first column shows the query sketches, each row shows top 10 results of retrieving, and the green box is true while the red is false. (a) Retrieval results on UCM. (b) Retrieval results on RS19.

result is true while the red one indicates it is false. Obviously, the proposed method provides satisfactory retrieval performance. The query sketches contain multiple dominant objects are considered. As showed in Fig.6(a), we draw a sketch which contains multiple dominant objects to represent golf course, and the proposed method obtains a good retrieval result. Moreover, sketches with single dominant objects can search out a complex scene with multiple dominant objects. For instance, when we depict an airplane sketch to search for airports, the scenes with several airplanes are found out. Besides, since the geometric construction of sea wave is similar to the riverbank in the UCM dataset, the river cannot be retrieved well.

Table 1 shows the mean average precision (MAP) and the precision of top K retrieval results while employing different features to retrieve aerial images in the UCM dataset with semantic sketches. The last two rows show the results of our single-scale model and multi-scale model trained by Aerial-SI. We find our multi-scale deep model provides the best results in terms of different metrics, and the gap between our model and the baselines is large. The results of using GoogLeNet are inferior to those of Alexnet. This phenomenon can be attributed to that the size of the dataset is too small to retrain the deep network as GoogLeNet. But using GooLeNet in our method still performs much better than using handcrafted features. The retrieval model is needed to be trained on one dataset and used on another unknown dataset. Therefore, we use the dataset Aerial-SI as the training set, and obtain the proposed model. Then we test the model on completely independent testing datasets. That is, categories in two datasets are different and aerial images are non-repeated, which means that the training dataset is irrelevant to the testing dataset. Hence, experiments demonstrate that the trained model is efficient on other unrelated datasets. It is proved that our multi-scale deep model can obtain features that represent both aerial images and sketches well.

Table 1. Comparative Results with Baselines

Feature	MAP	Top-10	Top-50	Top-100
GIST	0.130 5	0.145 5	0.129 1	0.118 5
BoW	0.051 7	0.067 3	0.056 4	0.057 1
SIFT+SPM	0.047 7	0.041 8	0.041 5	0.040 2
GF-HOG	0.071 1	0.120 0	0.102 2	0.085 5
HED+GF-HOG	0.082 6	0.158 2	0.117 1	0.097 6
GoogLeNet	0.287 7	0.567 3	0.394 2	0.278 0
Ours(single)	0.445 5	0.707 3	0.548 7	0.440 7
Ours(multi)	**0.463 6**	**0.721 8**	**0.567 3**	**0.458 2**

4.4 Further Analysis on Different Layers and Modeling Strategies

When using a deep cross-domain model in the sketch-based aerial image retrieval task, the performance of features varies from layer to layer. The research in [37] shows that the 6th layer feature performs the best when retrieving natural images, but things seem to be changed in our experiments when retrieving aerial images with semantic sketches. In Fig.7, Preliminary indicates the preliminary model, Multi-pre represents the multi-scale preliminary model, and Multi-RS is the proposed model. FC6+FC7 represents employing the 6th and the 7th fully-connected layers as deep features, and FC7 denotes only using the 7th fully-connected layers. Fig.7 shows that the top K precision of the 6th and the 7th layer connected feature performs better when using preliminary models, but the 7th full-connected layer feature performs the best on both UCM and RS19. We can find in Fig.8 that compared with the connected feature of the 6th and the 7th layer, the MAP of eight classes in UCM and six classes in RS19 is higher when retrieving with the 7th layer feature. From the above, we find the 7th layer feature performs the best during retrieving aerial images semantic sketches.

We also compare the performance of different modeling strategies. Fig.7 and Fig.8 show the multi-scale strategy performs better on both UCM and RS19. Fine-tuning the multi-scale preliminary model using the Aerial-SI dataset brings a significant improvement by making the model adjust to the aerial images. We also observe that the multi-scale strategy is effective in this sketch-based image retrieval mission, since the consideration of various levels of details benefits bridging the gap between sketches and images. Fig.9 shows the visualized results of different layers in Alexnet and our multi-scale deep model. The visualization is achieved using the method in [38]. It is observed that the fully-connected layers of sketches in our model obtain a similar color to aerial images and keep the structure information at the same time. It is proved that our model provides better performance in extracting the common feature of aerial images and sketches. Besides, the visualization of each scale shows that the multi-scale strategy obtains various levels of detailed information,

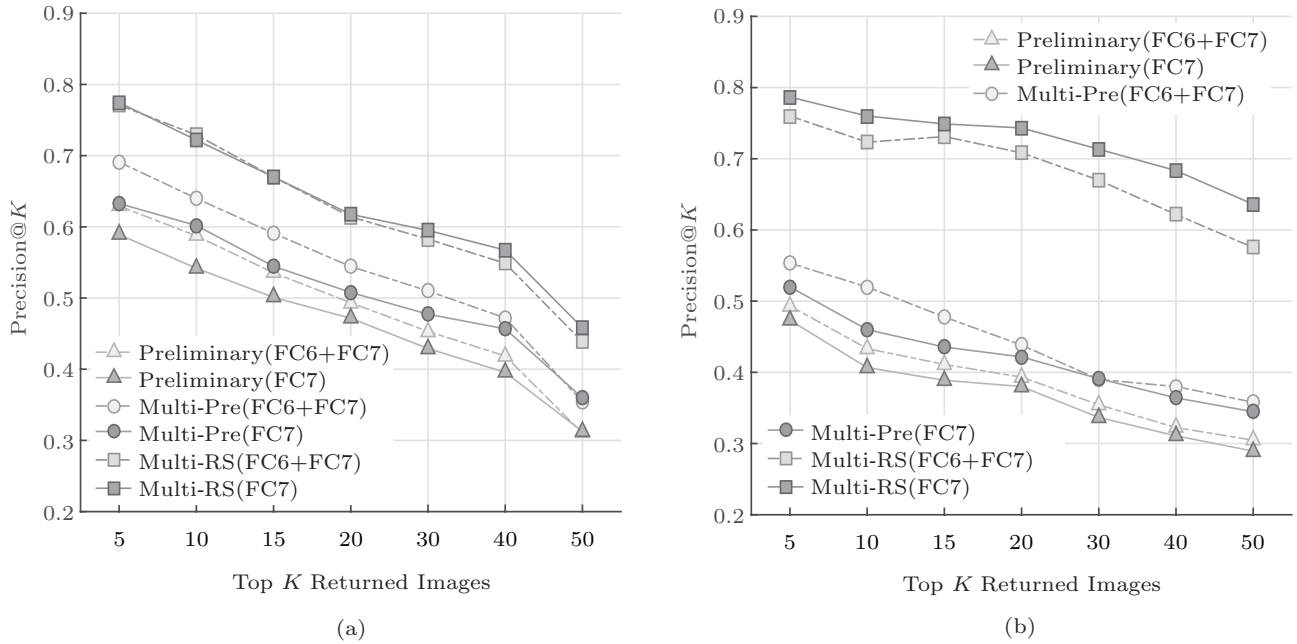


Fig.7. Top K precision of retrieving results on public datasets when using the proposed multi-scale deep model. The 7th fully-connected layers are used as deep features. (a) Top K precision on UCM. (b) Top K precision on RS19.

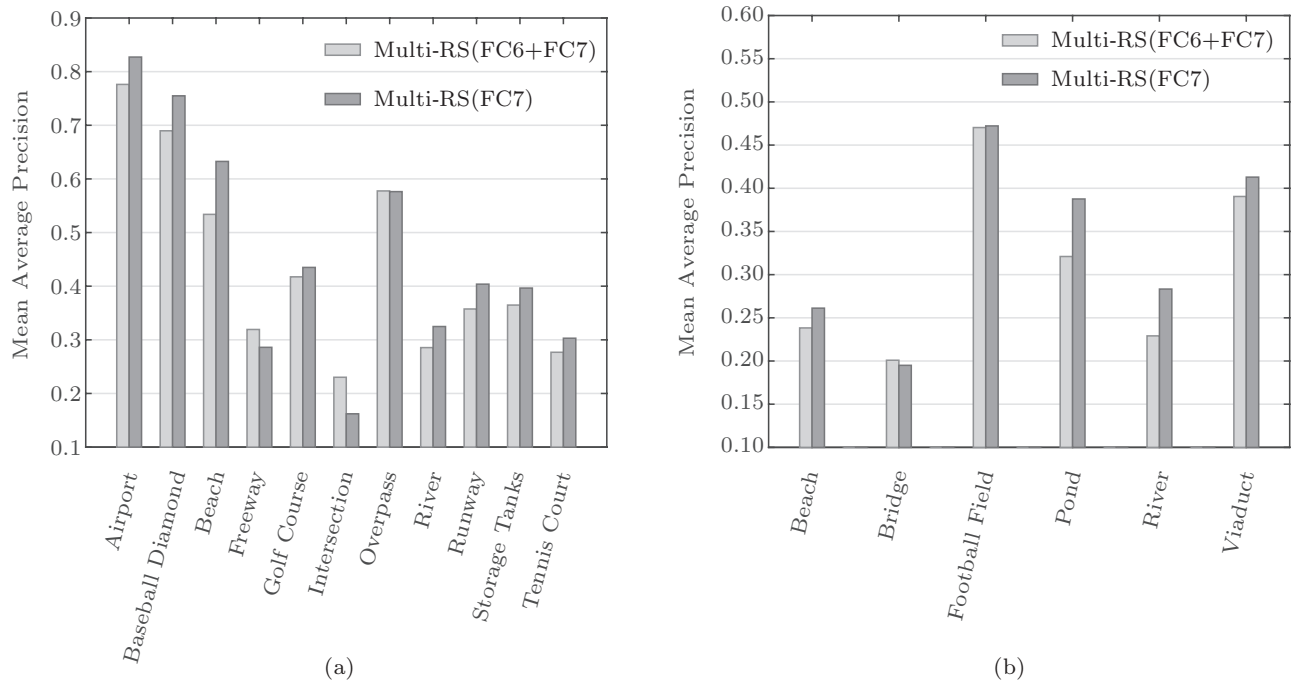


Fig.8. Mean average precision on the public datasets by using the proposed multi-scale deep model. The 7th fully-connected layers are used as deep features. (a) MAP on UCM. (b) MAP on RS19.

which means this strategy makes a contribution to the retrieval mission.

In the future, we will introduce metric learning to measure the similarity between cross-domain features, which may improve the retrieval results.

5 Conclusions

This paper proposed a multi-scale deep model to extract the cross-domain feature for retrieving aerial images with semantic sketches. Also, we contributed

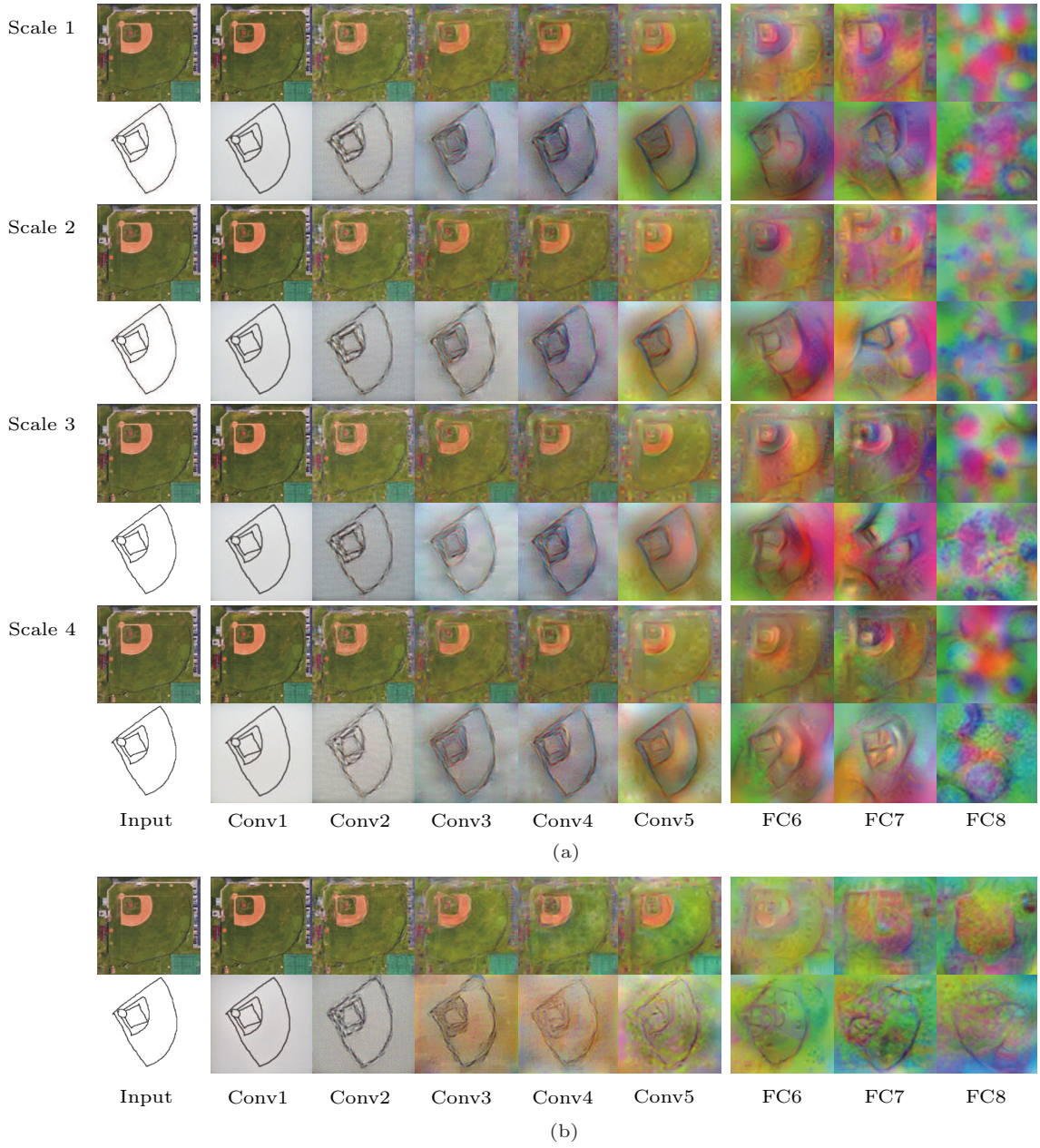


Fig.9. Reconstruction of the input data from conv1 to conv5, and FC6 to FC8. Conv: convolutional layer; FC: fully-connected layer. (a) Multi-scale deep model. (b) Alexnet.

an aerial sketch-image dataset Aerial-SI with annotation. The dataset was augmented to various levels of details to train a multi-scale deep model. Cross-domain features were extracted using the multi-scale deep model and then employed to measure the similarity between query sketch and aerial image. The experiments demonstrated the proposed method can overcome the cross-domain gap and the ambiguity inherent in sketches, and it obtains a good performance on the challenging sketch-based aerial image retrieval task. This method works well on independent aerial datasets.

In the future, we will introduce metric learning to measure the similarity between cross-domain features, which may improve the retrieval results.

References

[1] Hu F, Xia G S, Wang Z, Huang X, Zhang L, Sun H. Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification. *IEEE Journal of Selected Topics in Applied Earth Obser-*

vations and Remote Sensing, 2015, 8(5): 2015-2030.

[2] Xia G S, Wang Z, Xiong C, Zhang L. Accurate annotation of remote sensing images via active spectral clustering with little expert knowledge. *Remote Sensing*, 2015, 7(11): 15014-15045.

[3] Hu F, Xia G S, Hu J, Zhang L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, 2015, 7(11): 14680-14707.

[4] Aptoula E. Remote sensing image retrieval with global morphological texture descriptors. *IEEE Transactions on Geoscience and Remote Sensing*, 2014, 52(5): 3023-3034.

[5] Demir B, Bruzzone L. A novel active learning method in relevance feedback for content-based remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2015, 53(5): 2323-2334.

[6] Demir B, Bruzzone L. Hashing-based scalable remote sensing image search and retrieval in large archives. *IEEE Transactions on Geoscience and Remote Sensing*, 2016, 54(2): 892-904.

[7] Ozkan S, Ates T, Tola E, Soysal M *et al*. Performance analysis of state-of-the-art representation methods for geographical image retrieval and categorization. *IEEE Geoscience and Remote Sensing Letters*, 2014, 11(11): 1996-2000.

[8] Ferecatu M, Boujemaa N. Interactive remote-sensing image retrieval using active relevance feedback. *IEEE Transactions on Geoscience Remote Sensing*, 2007, 45(4): 818-826.

[9] Du Z, Li X, Lu X. Local structure learning in high resolution remote sensing image retrieval. *Neurocomputing*, 2016, 207: 813-822.

[10] Yang Y, Newsam S. Geographic image retrieval using local invariant features. *IEEE Transactions on Geoscience and Remote Sensing*, 2013, 51(2): 818-832.

[11] Liu T, Zhang L, Li P, Lin H. Remotely sensed image retrieval based on region-level semantic mining. *EURASIP Journal on Image Video Processing*, 2012, 2012(1): 4-15.

[12] Eitz M, Hays J, Alexa M. How do humans sketch objects? *ACM Transactions on Graphics*, 2012, 31(4): 44:1-44:10.

[13] Bai X, Latecki L J. Path similarity skeleton graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, 30(7): 1282-1292.

[14] Wang X, Feng B, Bai X, Liu W, Latecki L J. Bag of contour fragments for robust shape classification. *Pattern Recognition*, 2014, 47(6): 2116-2125.

[15] Bai X, Bai S, Zhu Z, Latecki L J. 3D shape matching via two layer coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(12): 2361-2373.

[16] Shen W, Wang X, Wang Y, Bai X, Zhang Z. DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), June 2015, pp.3982-3991.

[17] Eitz M, Hildebrand K, Boubekeur T, Alexa M. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE Transactions on Visualization Computer Graphics*, 2011, 17(11): 1624-1636.

[18] Hu R, Collomosse J. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision Image Understanding*, 2013, 117(7): 790-806.

[19] Cao Y, Wang C, Zhang L, Zhang L. Edgel index for large-scale sketch-based image search. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), June 2011, pp.761-768.

[20] Chen T, Cheng M M, Tan P, Shamir A, Hu S M. Sketch2Photo: Internet image montage. *ACM Transactions on Graphics*, 2009, 28(5): 124:1-124:10.

[21] Yu Q, Liu F, Song Y Z, Xiang T, Hospedales T M, Loy C C. Sketch me that shoe. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), June 2016, pp.799-807.

[22] Qi Y, Song Y Z, Zhang H, Liu J. Sketch-based image retrieval via Siamese convolutional neural network. In *Proc. IEEE International Conference on Image Processing* (*ICIP*), September 2016, pp.2460-2464.

[23] Wang X, Duan X, Bai X. Deep sketch feature for cross-domain image retrieval. *Neurocomputing*, 2016, 207: 387-397.

[24] Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification. In *Proc. the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, November 2010, pp.270-279.

[25] Xia G S, Yang W, Delon J, Gousseau Y, Sun H, Maître H. Structural high-resolution satellite image indexing. In *Proc. ISPRS TC VII Symposium-100 Years ISPRS*, Volume 38, September 2010, pp.298-303.

[26] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems* (*NIPS*), December 2012, pp.1097-1105.

[27] Zhang Q, Shen X, Xu L, Jia J. Rolling guidance filter. In *Proc. the 13th European Conference on Computer Vision* (*ECCV*), September 2014, pp.815-830.

[28] Hu J, Jiang T, Tong X, Xia G S, Zhang L. A benchmark for scene classification of high spatial resolution remote sensing imagery. In *Proc. IEEE International Geoscience and Remote Sensing Symposium* (*IGARSS*), July 2015, pp.5003-5006.

[29] Xia G S, Hu J, Hu F, Shi B, Bai X, Zhong Y, Zhang L. AID: A benchmark dataset for performance evaluation of aerial scene classification. *arXiv:1608.05167*, 2016. https://arxiv.org/abs/1608.05167v1, May 2017.

[30] Zitnick C L, Dollár P. Edge boxes: Locating object proposals from edges. In *Proc. European Conference on Computer Vision* (*ECCV*), September 2014, pp.391-405.

[31] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 2011, 42(3): 145-175.

[32] Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, 60(2): 91-110.

[33] Grauman K, Darrell T. The pyramid match kernel: Discriminative classification with sets of image features. In *Proc. the 10th IEEE International Conference on Computer Vision* (*ICCV*), October 2005, pp.1458-1465.

[34] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), June 2006, pp.2169-2178.

[35] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), June 2005, pp.886-893.

[36] Xie S, Tu Z. Holistically-nested edge detection. In *Proc. IEEE International Conference on Computer Vision* (*ICCV*), December 2015, pp.1395-1403.

[37] Babenko A, Slesarev A, Chigorin A, Lempitsky V. Neural codes for image retrieval. In *Proc. European Conference on Computer Vision* (*ECCV*), September 2014, pp.584-599.

[38] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), June 2015, pp.5188-5196.

Tian-Bi Jiang received her B.S. degree in remote sensing science and technology from Wuhan University, Wuhan, in 2015. She is currently pursuing her M.S. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan. Her current research mainly focuses on sketch-based image retrieval and image understanding.

Gui-Song Xia received his B.S. degree in electronic engineering and M.S. degree in signal processing from Wuhan University, Wuhan, in 2005 and 2007, respectively, and his Ph.D. degree in image processing and computer vision from the CNRS LTCI, TELECOM ParisTech, Paris, in 2011. Since March 2011, he has been a postdoctoral researcher with the Centre de Recherche en Mathmatiques de la Decision (CEREMADE), CNRS, Paris-Dauphine University, Paris, France, for one and a half years. Currently, he is a professor with the State Key Laboratory of Information Engineering, Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University, Wuhan. His research interests include mathematical image modeling, texture synthesis, image indexing and content-based retrieval, structures from motions, perceptual grouping, and remote-sensing imaging.

Qi-Kai Lu received his Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, in 2016. He is currently with the School of Electronic Information, Wuhan University. His research interests include image processing, machine learning, and remote sensing applications.

Wei-Ming Shen received his Ph.D. degree in mathematics from Wuhan University, Wuhan, in 1993. He is currently a professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan. His research interests include video analysis and image processing.